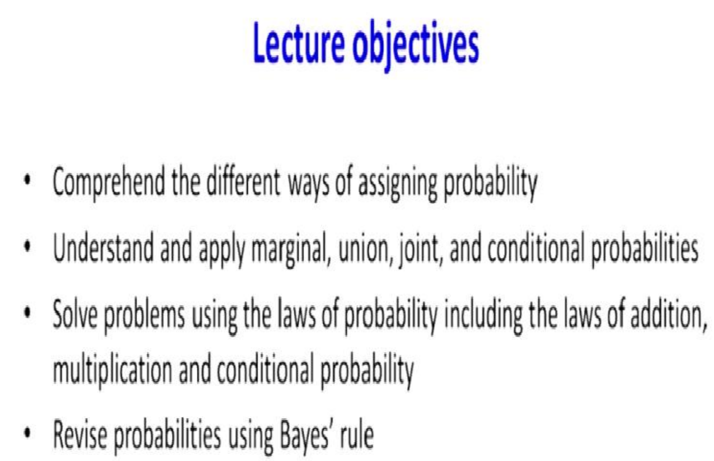
Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture – 06
Introduction to Probability - I

Good morning students. Today we are going to next lecture number 6 introduction to probability. The concept of probability is fundamental telefield whether you call it a statistics or analytics are your data science everywhere because if you look at some of the book titles of the statistics or analytics it will come with probability and statistics because the concept of probability and statistics is cannot separate it because always will go together because, the concept of statistics since we are taking sampling, we are predicting about the population.

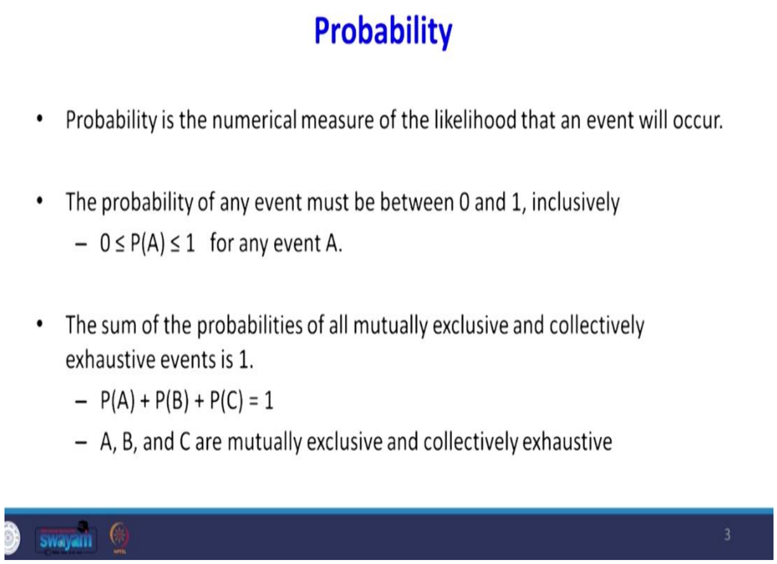
So, whatever we say with the help of sampling we have to attach always some probability because it cannot be 100% assured that whatever you say with the help of sample will be exactly; you cannot predicted so when there is a prediction comes, then you have to attached probability to that. Today, you will see that it is an introduction to probability I am not going to teach in detail about that one.

(Refer Slide Time: 01:21) correct the picture size



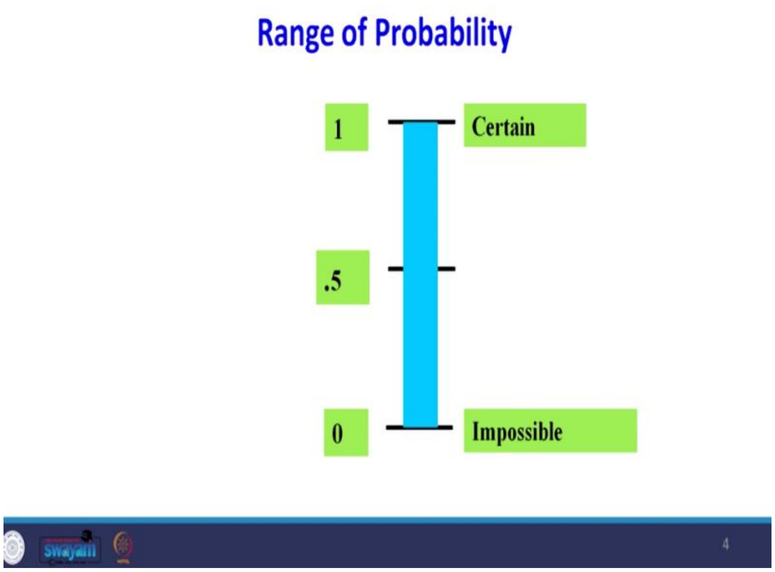
What are the ideas which are important for us the only that I am going to teach. So the lecture objective is to comprehend the different way of assigning probability understand and apply in marginal union joint and conditional probabilities and solving problem using laws of probability including law of addition, multiplication and conditional probability and using very important theorem that is the Bayes rule that to revise the probability at the end these are the my lecturer objectives.

(Refer Slide Time: 01:51)



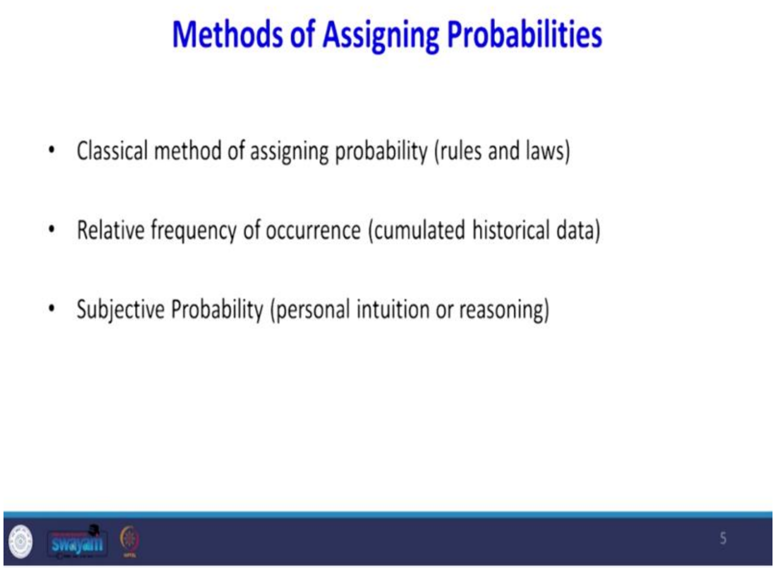
So we will go the definition of the probability, the probabilities the numerical measure of likelihood that an event will occur, the probability of any event must be been between 0 and 1, inclusively 0 to 1 for any event A, the sum of the probability of all mutually exclusive collectively exhaustive event is 1 latter will explain what would be mutually exclusive collectively exhaustive events. So always the probability of summation of probability will be equal to 1, for example, probability A plus probability B and probability C equal to 1, here A, B and C are mutually exclusive and collective events.

(Refer Slide Time: 02:32)



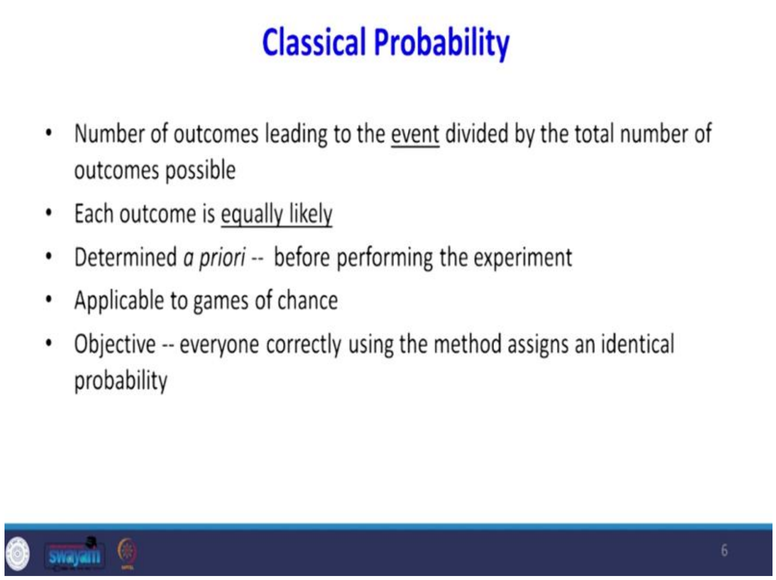
So this is the range of probability you see that, if it is an impossible event, the probability value is 0. If it is a certain event, the probability is 1 if it is there a 50% chance, the probability is point 5. So, the point here is it is the probability lies between 0 to 1.

(Refer Slide Time: 02:50)



The method of assigning probability there are 3 methods, one is the classical method of assigning probability rules and laws, the relativity frequency of occurrence that is cumulative to historical data and subject to probability that is a personal intuition or reasoning by using these 3 methods, let us see how to find out the probability.

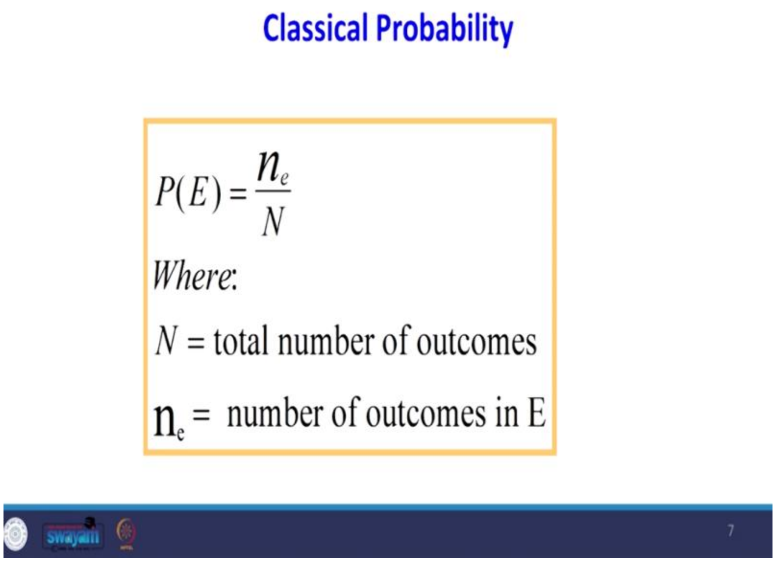
(Refer Slide Time: 03:11)



First of all is a classical probability the number of outcomes leading to an event divided by the total number of outcomes possible is a classical probability. Each outcome is equally likely there is an equal chance of getting different outcome. So, it is determined a priori that is before performing the experiment we know what are the outcome is going to come suppose we toss a coin there are 2 possibility head or tail in advance we know what are the possible outcomes.

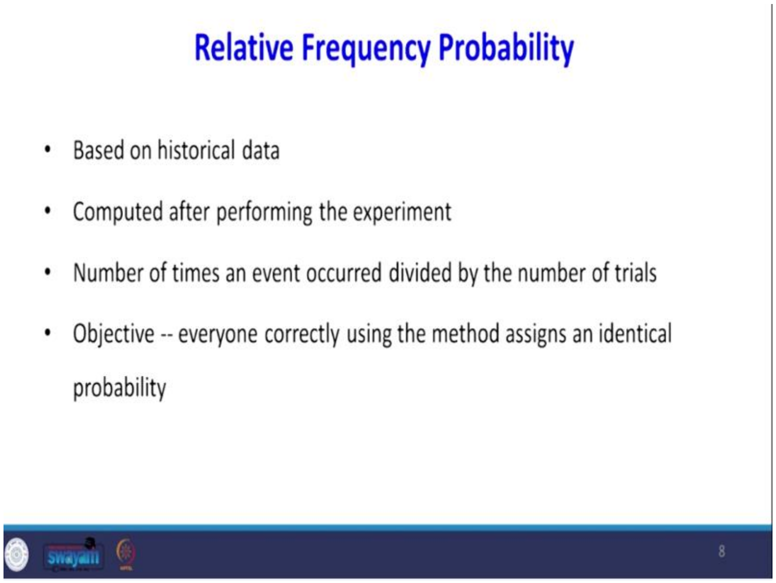
It is applicable to games of chance the object to is everyone correctly using the method assign and identical probability because what is happening here, using classical probability that everyone will get the same answer for a problem because we know in advance what are the possible outcomes.

(Refer Slide Time: 04:00)



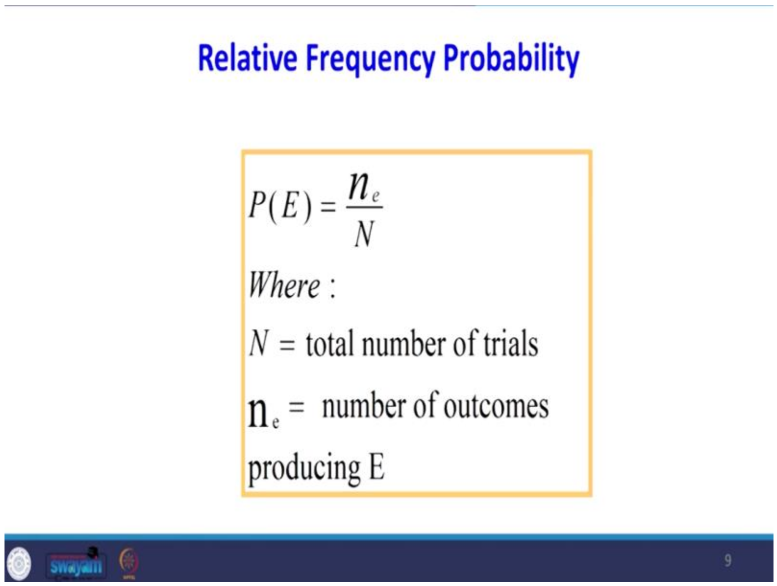
So, mathematically the probability of E is n of e divided by capital N where the capital N is the total number of outcomes n of e is the number of outcomes in E.

(Refer Slide Time: 04:15)



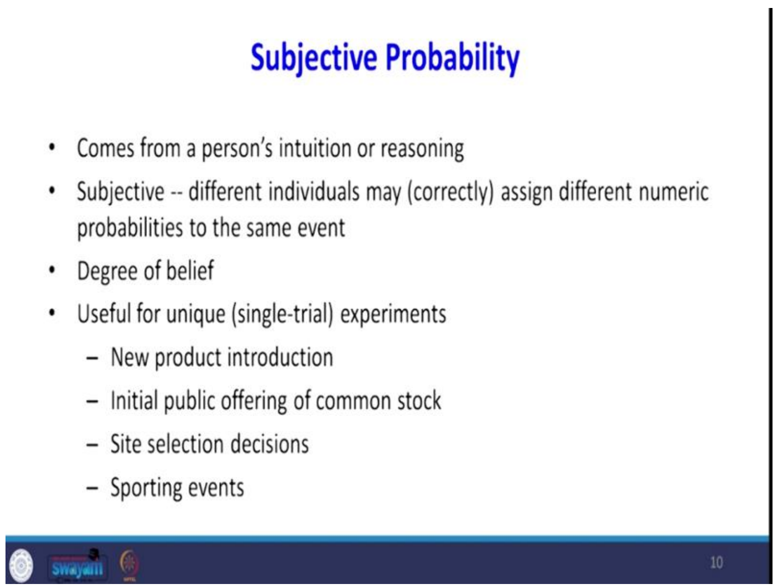
The relative frequency probability, it is based on the historical data because the another name for relatively frequencies of probability, it is computed after performing the experiment, number of items an event occurred divided by number of trials and the but frequency divided by some frequency, here also everyone correctly using this method assign as identical probability because everything is already known to you.

(Refer Slide Time: 04:42)



So, the same formula probability E is n of e divided by N where n e is number of outcomes N is total number of trials.

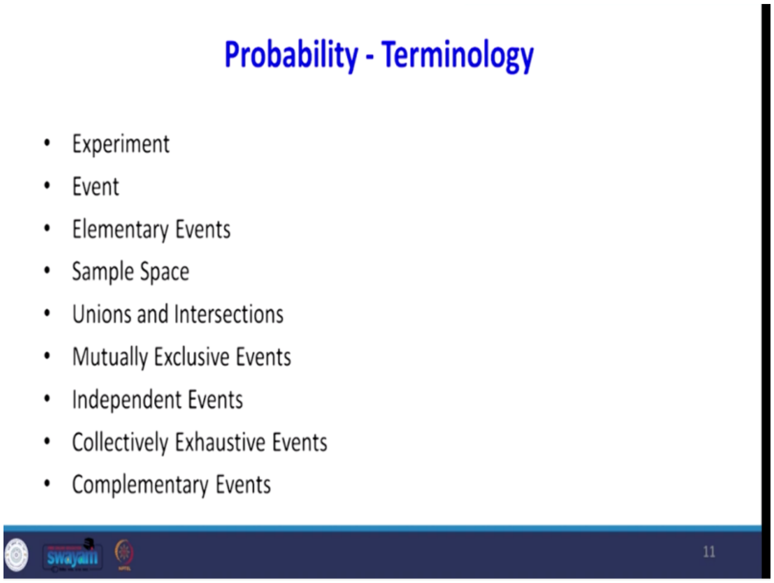
(Refer Slide Time: 04:50)



Then subject to probability, it comes from a person's intuitions or reasoning. Subjective means different individuals may correctly assign different numerical probabilities to the same event, it is the degree of belief sometimes subjective probability is useful for example, if you introduce a new product, suppose if you want to know the probability of success of the new product so we can ask an expert that what is the probability of success even it is a new movie or new project what is the probability of success?

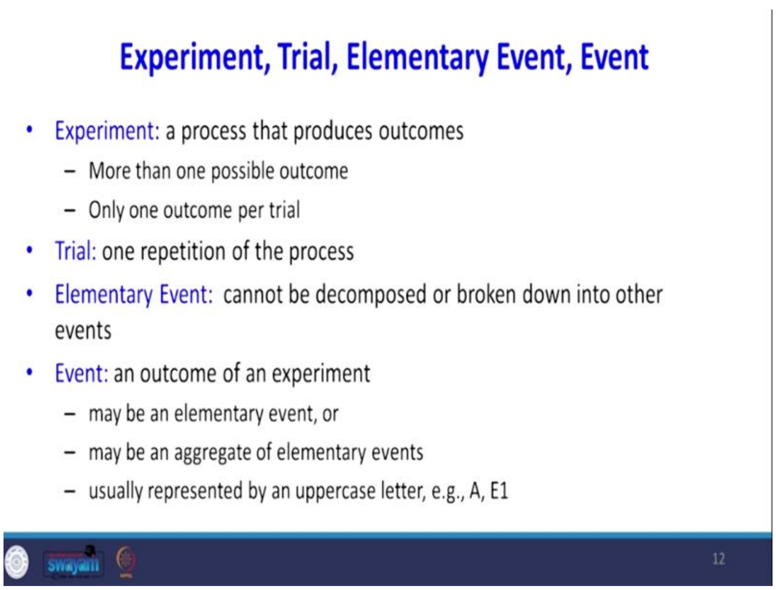
It is based on the intuition are based on the experience of the person he can give some probability of success or failure for example sites select decision for sporting events for example, in cricket what is the how much possibility of one team to win these are intuitive probability.

(Refer Slide Time: 05:42)



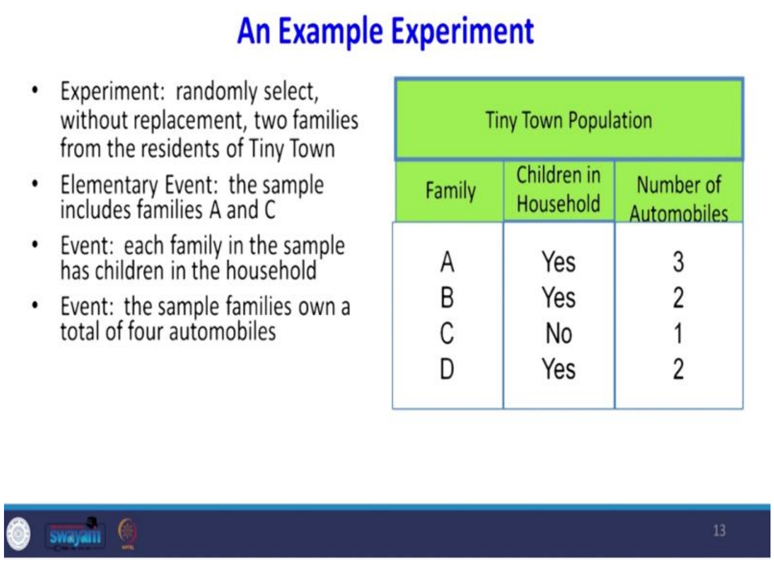
In this course that there are certain terminology with respect to probability you have to know it is very fundamental, even though you might have studied or even with previous classes it is just to recollect it what is the experiment we will see what is the experiment event, elementary event, sample space, union and intersections mutually exclusive events, independent events, collectively exhaustive events, complimentary events, these are some terms we will revise this.

(Refer Slide Time: 06:07)



One is we say what is experiment trial elementary event and event experiment, a process that produces outcomes is experiment so there are more than one possible outcome is there only one outcome per trial, what is a trial one repetition of process is a trial what is the elementary event? Even that cannot be decomposed or broken down into other events that is elementary events and what is the event and outcome of an experiment may be an elementary event maybe aggregate of elementary event usually represented by uppercase letter for example A, E that is notation for event.

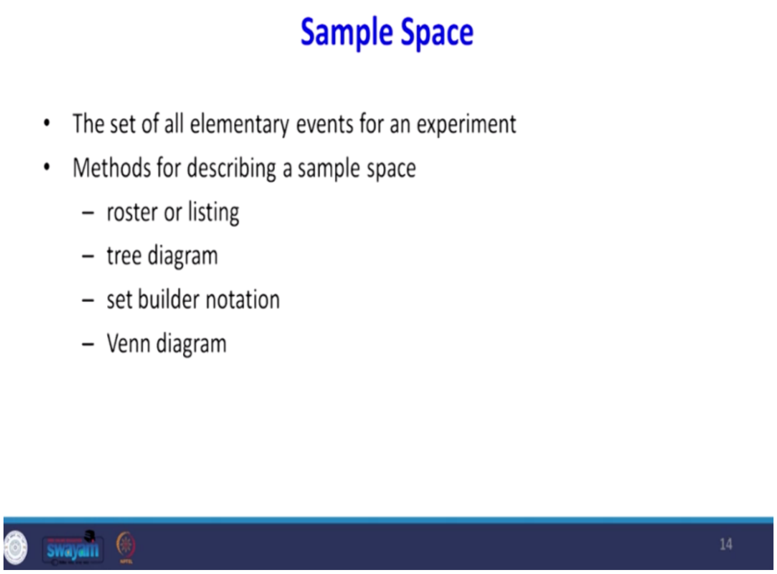
(Refer Slide Time: 06:48)



We look at this example. If you look at the table there are some towns population is given there are 4 families their family A, B, C, D we asked the 2 questions, children's in household whether do you have children are not see family yes then we asked a number of automobiles how many number of automobiles you have 3. B they have children, they have 2 automobiles for the help of this table will try to understand what is experiment for example, randomly select without replacement 2 families from the residents of the town.

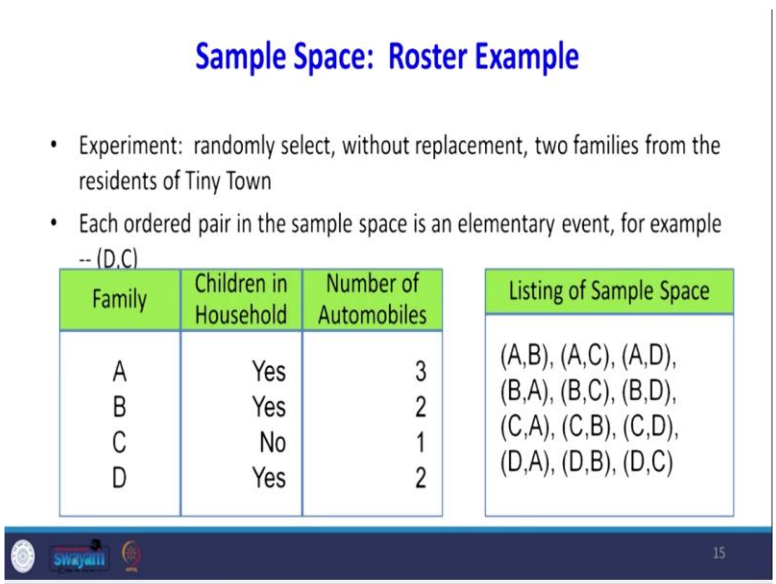
For example, randomly we can select so elementary event for example, the sample includes family A and C randomly you have to selected. So, event each family in the sample has children in the household, the sample families own a total of 4 automobiles to these are particular events for example for the event each family in the sample has children in the household for example A is one event D is another event for example, the sample families own a total number of 4 automobiles for example, A and C they have 4 automobiles B and D they have 4 automobiles A and D they have more than 4 automobiles.

(Refer Slide Time: 08:07)



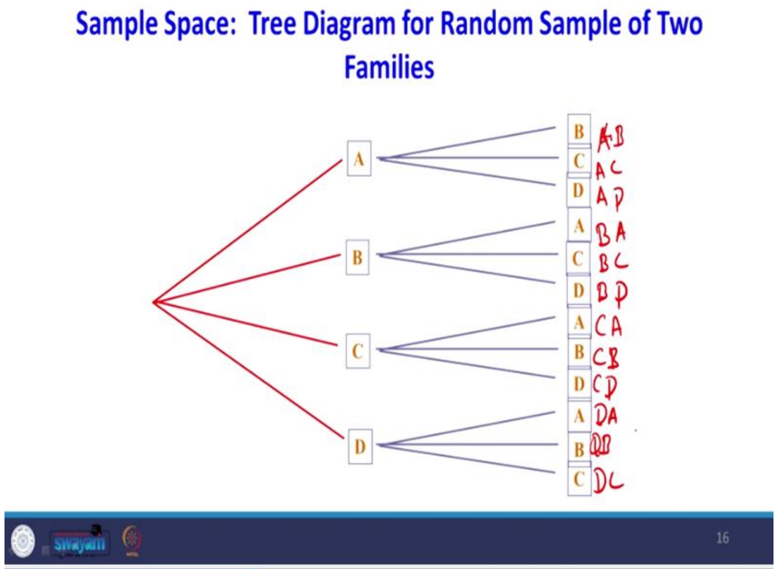
This is the example of event then what is the sample space, the set of all elementary events for an experiment is called a sample space. Suppose if you roll a die, there are you can get 1 2 3 4 5 6 these are the sample space there are different methods for describing the sample space one is listing tree diagram, set builder notation and Venn diagram, you will see what is that.

(Refer Slide Time: 08:30)



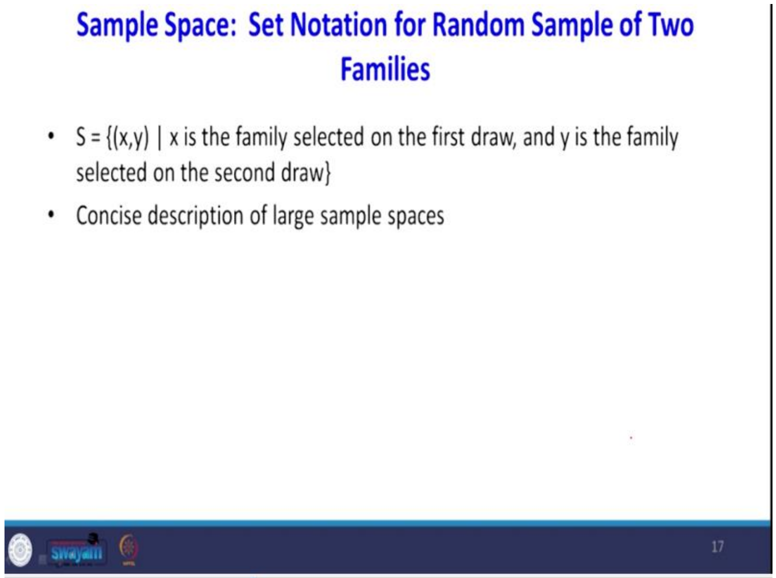
See this listing experiment randomly select without replacement 2 families from the residents of the town, so each order the pair in the sample spaces elementary event for example D, C. So what are the different possibility look at this table A B, A C, A D, B A, B C, B D, C A, C B, C D so, these are the listing the sample space what is that we have to select 2 families from the residents.

(Refer Slide Time: 08:58)



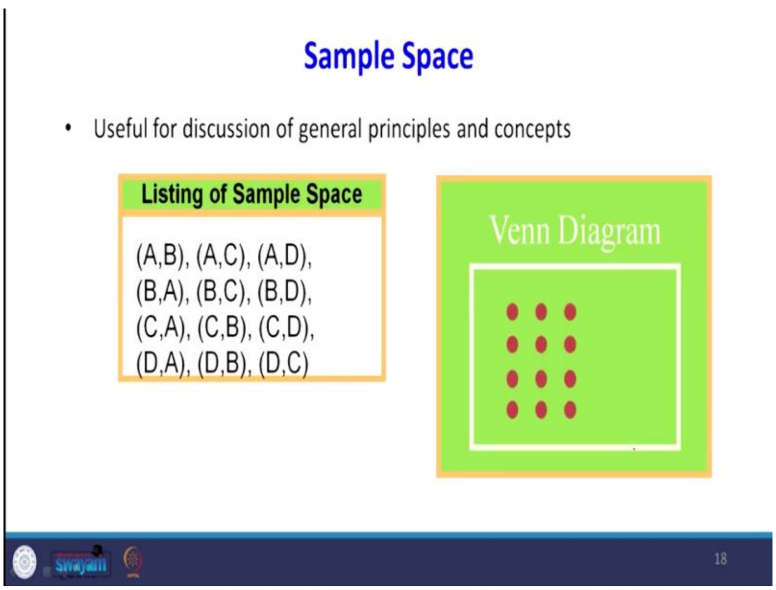
So, here we will do without replacement, without replacement means suppose A it owned by again A, if it is B it owned by again B, once A is taken we are not selecting another A, so without replacement the same thing the another way to express the sample spaces with the help of tree diagram, it is a tree diagram is very useful and easy to understand. For example A B C D there are 4 families we can have combination A B we can combination A C, A D, B A, B C, B D, C A, C B,C D, D A, D B, D C. So, this is the easy which is the different sample space because tree diagram is easy to understand.

(Refer Slide Time: 09:55)



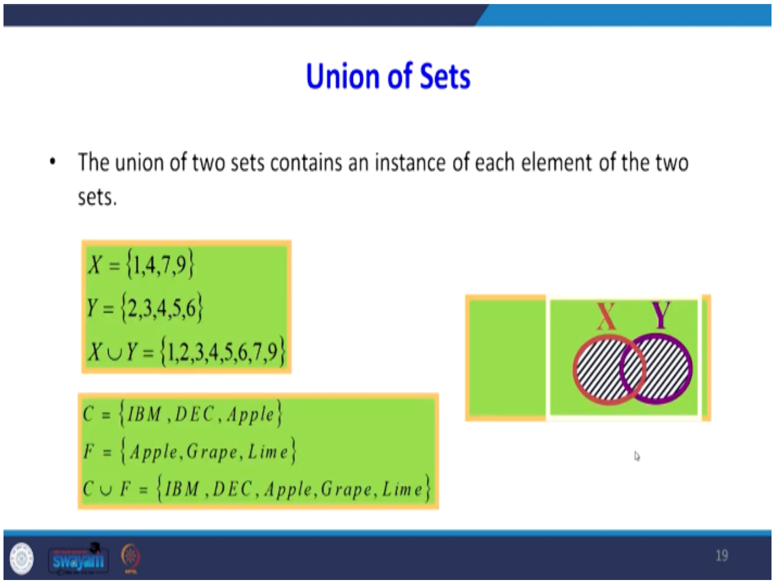
Now the set notation for random sample of 2 families so S = {(X, Y) X is the family selected on the first draw, and Y is the family selected on the second draw}. It is the concise description of larger sample spaces in mathematics they use this kind of notations.

(Refer Slide Time: 10:15)



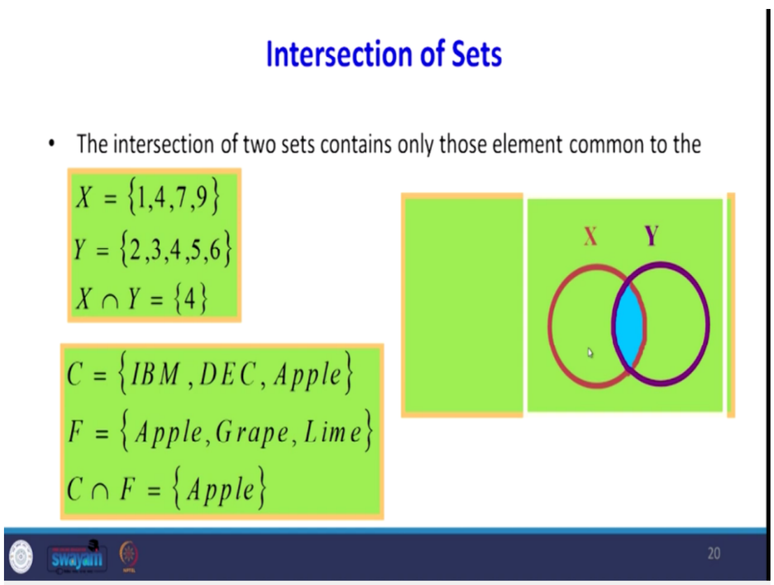
You see the sample space can be shown in terms of Venn diagram, so this is a list of sample space see that this is a different dot express different sample space

(Refer Slide Time: 10:26)



Then we will go to the another concept union of sets, the union of 2 sets contains an instance of each element of the 2 sets for example X is 1,4,7,9 one set, Y is another set 2,3,4,5,6 So, if you want to know X union Y just we have to combine 1,2,3,4,5,6,7,9 similarly we look at the Venn diagram X is 1 Y is 1 if you want to know union combining both events and other examples say C IBM, DEC, Apple that is the C set there is the another set F Apple, Grape, Lime suppose we want to know union of set C and F so we are to take IBM, DEC, Apple, Apple is coming in both sets we are taking only one grape lime this is the union.

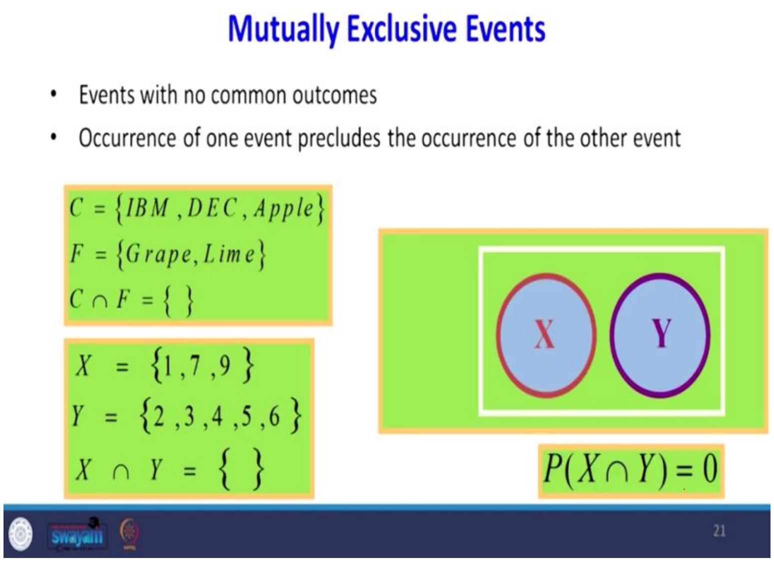
(Refer Slide Time: 10:15)



We go for intersection suppose for X = 1, 4, 7, 9 Y is 2, 3, 4, 5, 6 if you want to know common intersect is 2 sets contain only those elements common so here the forest common in X and Y so

X intersection Y is 4, for example, C and F, in C we have IBM, DEC, APPLE F is APPLE, GRAPE, LIME then C intersection F that is a common thing between set C and F is Apple, so this one see this portion says our intersection.

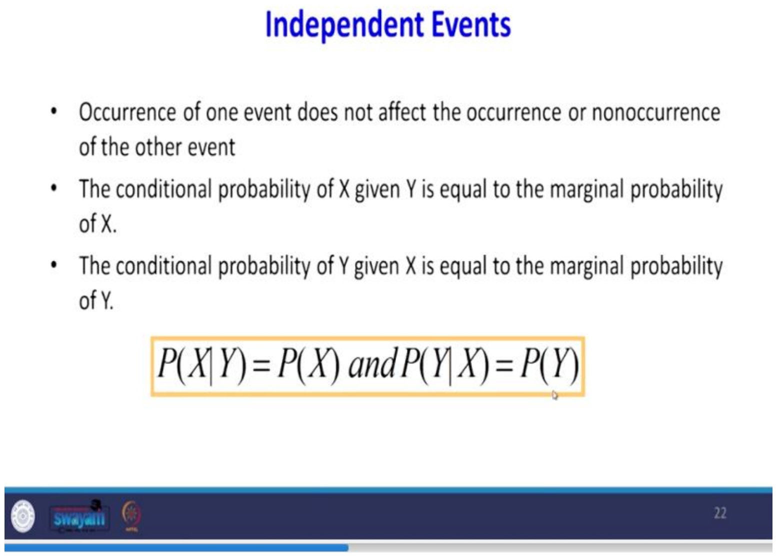
(Refer Slide Time: 11:54)



Then we will go for mutually exclusive events even with the no common outcomes is called mutually exclusive events occurrence of one event precludes the occurrence of other event for example C IBM, DEC, Apple F is Grape, Lime. So C intersection there is no common thing so that is why it is null set similarly X is 1,7,9 Y is 2,3,4,5,6 X intersection Y there is no common set look at the Venn diagram there is no common thing so X intersection Y 0 these 2 sets are not over lapping.

So it is called mutual exclusive events. another example for this when we toss a coin there is 2 possibility to get the outcome one may be head or tail it cannot have both that is why both events are mutually exclusive event.

(Refer Slide Time: 12:46)

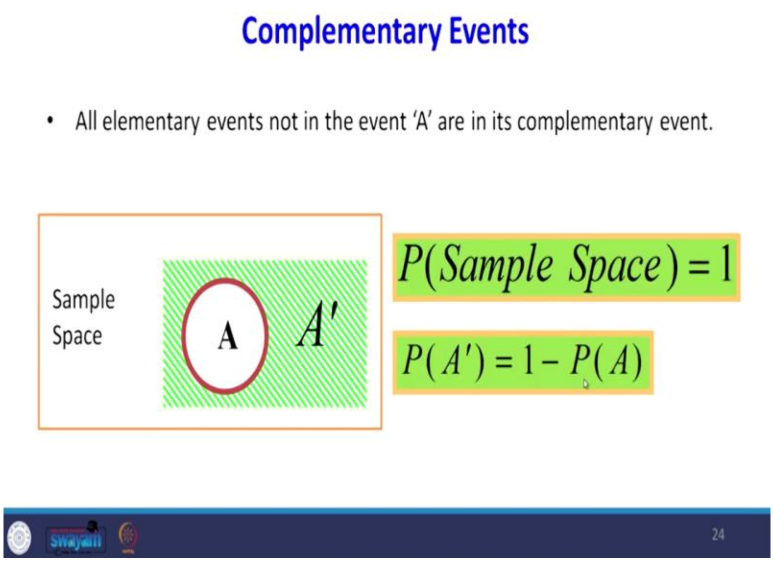


Then independent events, so occurrence of one event does not affect the occurrence or non occurrence of other event is called independent event, the conditional probability of X given Y is equal to the marginal probability of X the conditional probability of Y given X is equal to the marginal probability the one way we will do a small problem on this one way to test the independent event is suppose the P of X given Y equal to P X and P of Y given X is P of Y then even X and Y are called independent events will go in detail after some time but with the help of an example.

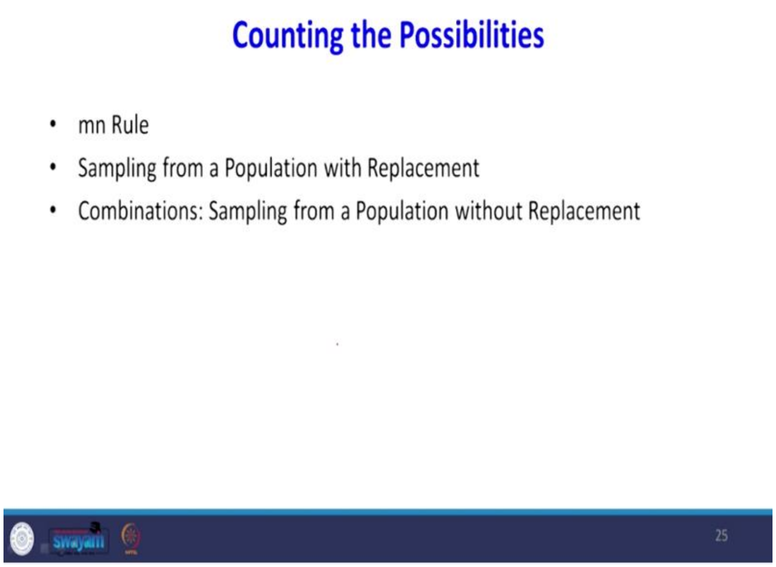
(Refer Slide Time: 13:28)

Collectively exhaustive event it contains all elementary events for an experiment suppose E1 E2 E3 sample space with 3 collectively exhaustive event suppose you roll your die, all possible outcome 1, 2, 3, 4, 5, 6 that is collectively exhaustive events.

(Refer Slide Time: 13:52)

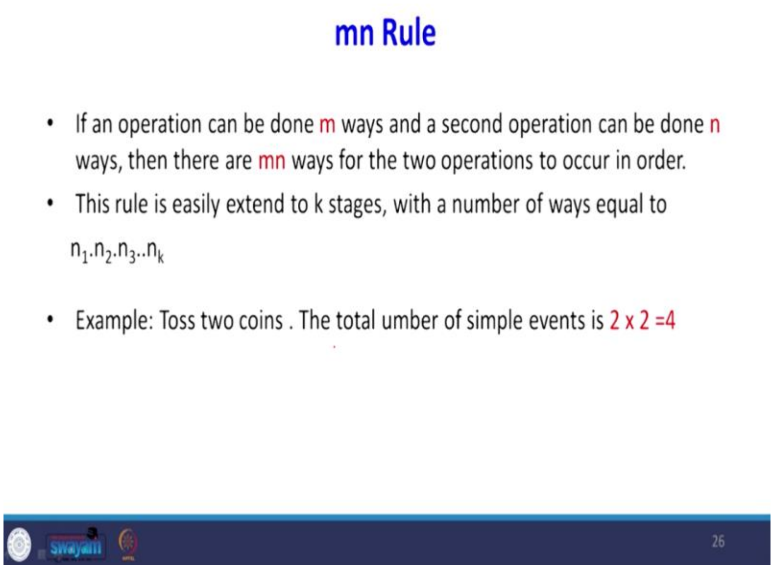


(Refer Slide Time: 14:11)



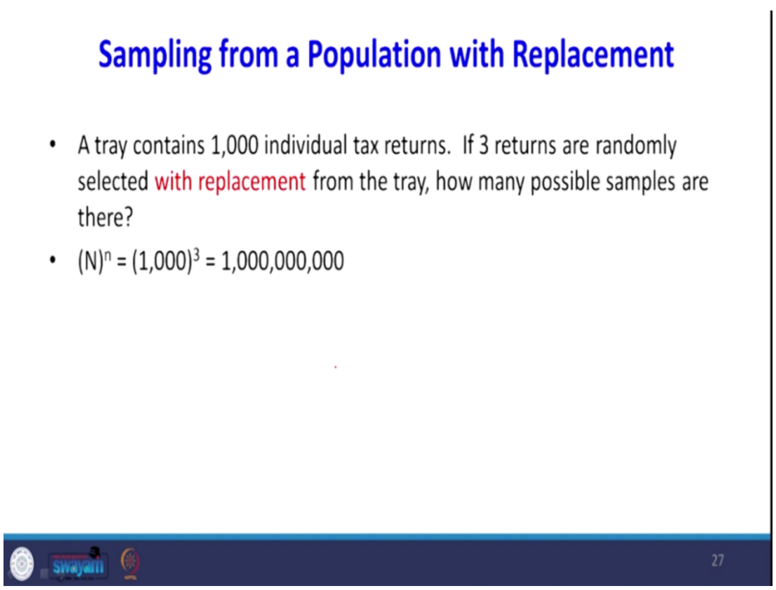
Then complementary events and elementary events not in the A dash or is it is complimentary event you see that the P of A is there is there which is not there that is A dash that is called complimentary, so P A dash is equal to 1 – P of A then counting the possibilities because in probabilities many time different combinations may come these rules may be very useful for counting different possibilities one rule is mn rule second one is sampling from a population with replacements, second one is sampling from a population without replacement.

(Refer Slide Time: 14:28)



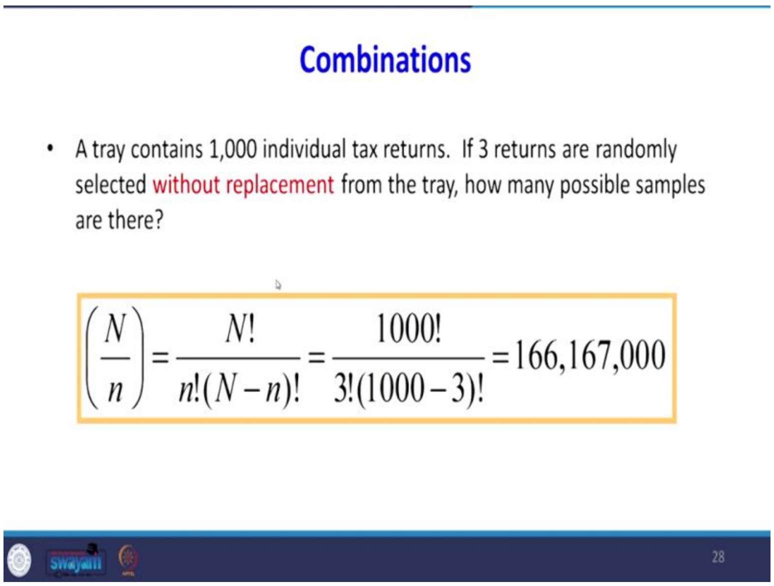
Will go for the mn Rule if an operation can be done m ways and the second operation can be done n ways, then there are mn ways for the 2 operation to occur in order. The rule is easily can be extended to k stages, with a number of ways equal to if there are k stages n1, n2, n3 there some simply we have to multiply for example toss 2 coins the total number of sample event is 2 multiply by 2 equal to 4 because in the first coin you make a 2 possibilities second coin you make it another 2 possibilities so the total is 4 possibilities.

(Refer Slide Time: 15:07)



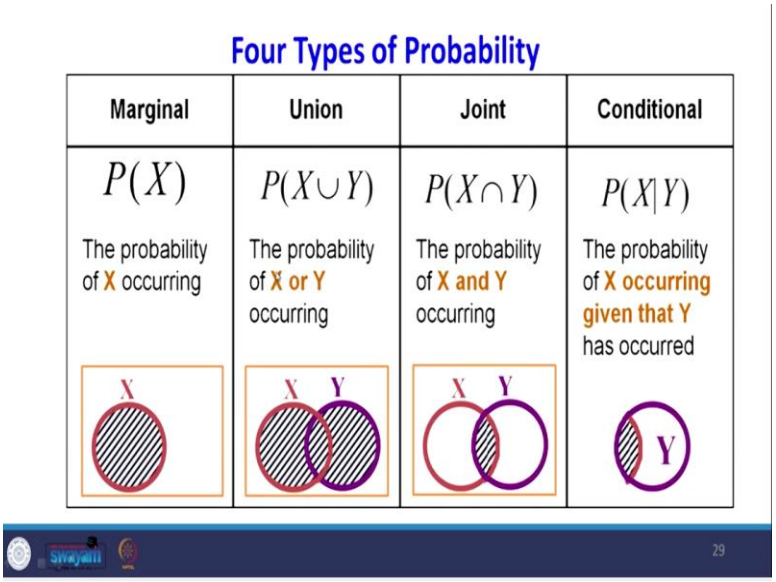
Suppose you see that another example of sampling from a population with replacement. One example is a tray contains 1000 individual tax returns if 3 returns are randomly selected with replacement from the tray, how many possible samples are there? So every time you are going for a 3 trial; trial 1 trial 2 trial 3, in each trial there are 1000 possibilities because you can choose one from 1000, so, firstly trial 1000 second trial 1000 third trial 1000 when you multiply this is 1000 million possibilities are there with replacement.

(Refer Slide Time: 15:45)



In case if you go without replacement, the same thing because without replacement what will happen the sample size will decrease. A tray contains 1000 individual tax returns in 3 returns are randomly selected without replacement from the tray. How many possible samples are there So, that is a N see n that is N factorial divided by n factorial multiplied by N – n factorial that is 1000 factorial divided by 3 1000 – 3 factorial is 166,167,000 you see the previously with replacement and going to previous light with replacement it is a 1000 million now, it is only 166 million because we are going for without replacement.

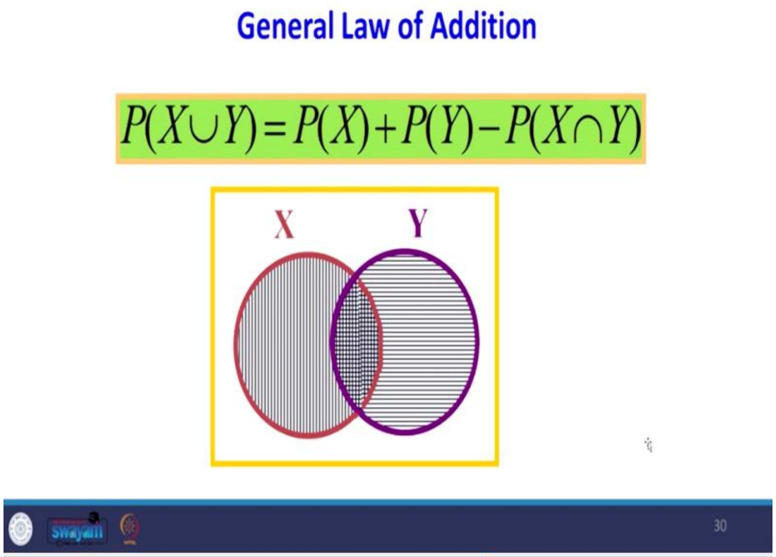
(Refer Slide Time: 16:29)



There are different types of probability say we can call it as a marginal probability union probability joint probability and conditional probability. Then what is the rotation model probabilities simple one probability P of X, so, how it is expressed in terms of Venn diagram, see this one, so marginal probability the union probability is the X union Y, the probability of X or Y counting's, the joint probability or common probability, the probability of X and Y occurring together the middle portions.

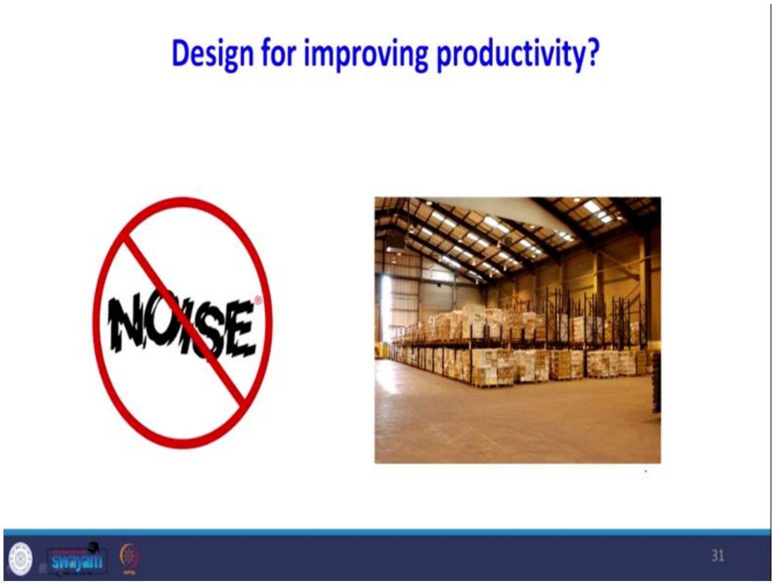
Then conditional probability, the probability of X occurring given that Y has occurred here there are 2 events, the probability of the outcome of X is depending upon the outcome of Y. So we have to read the probability of X given that Y has occurred. So this is the Venn diagram notation for expressing the conditional probability

(Refer Slide Time: 17:23)



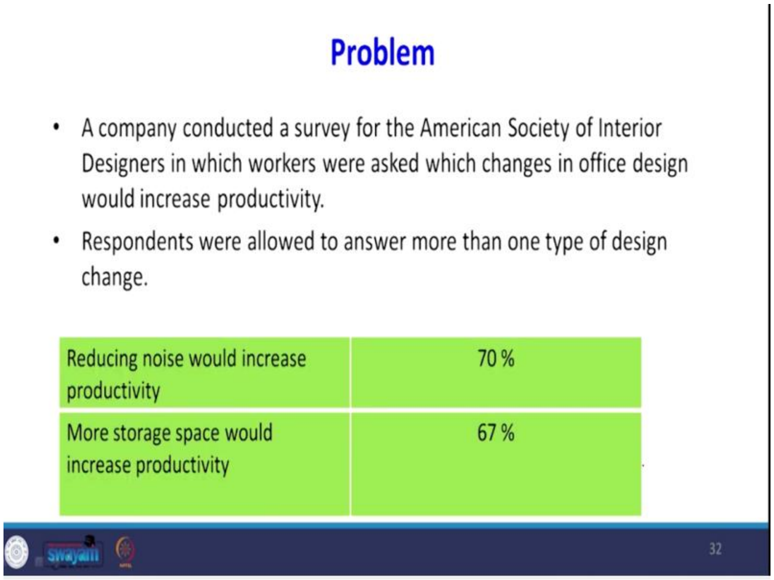
Then we will go for general law of addition so, P of X union Y equal to P of X + P of Y - P of X intersection Y.

(Refer Slide Time: 17:35)



Will take a small example, from that example, will understand the concept of probabilities a company is going for improving the productivity of the particular unit. They are coming with a new design one design is layout design for example, layout design, one design will reduce the noise, that is one option that is second design that will give you more storage space, so we are going to ask from the employees what kind out of these 2 designs which design will improve the productivity.

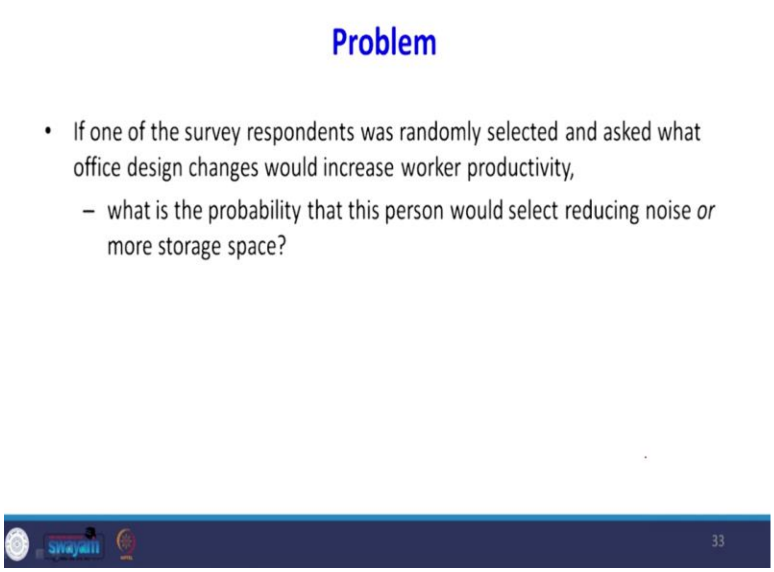
(Refer Slide Time: 18:13)



You see the problem, a company conducted a survey for the American Society of interior design in which workers were asked which changes in the office design would increase productivity, there are 2 design is there one is the one design will reduce the noise another design will improve the storage space. The responders were allowed to answer more than one type of design changes.

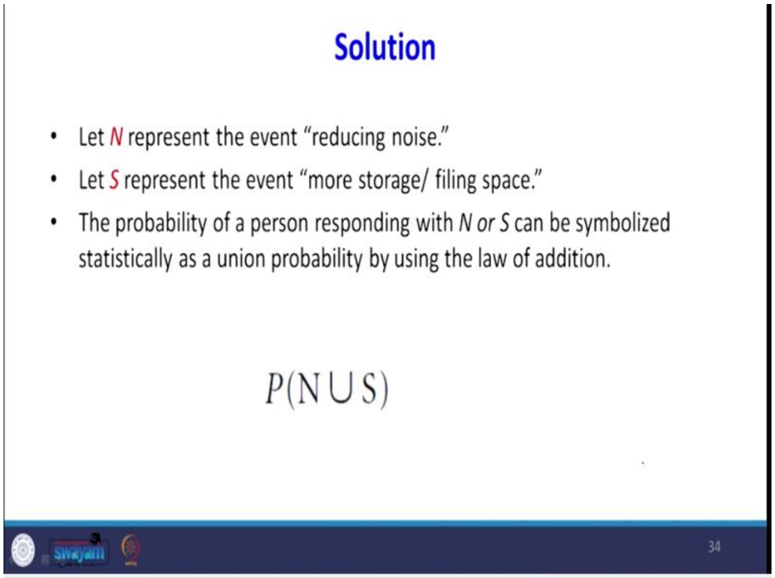
So this table shows the outcome so 70% as the people have responded that reducing noise would increase the productivity 67 percentage of the respondents responded that more storage space would increase the productivity there is the 2 design So, we are asking there from the respond which design will improve the productivity.

(Refer Slide Time: 19:04)



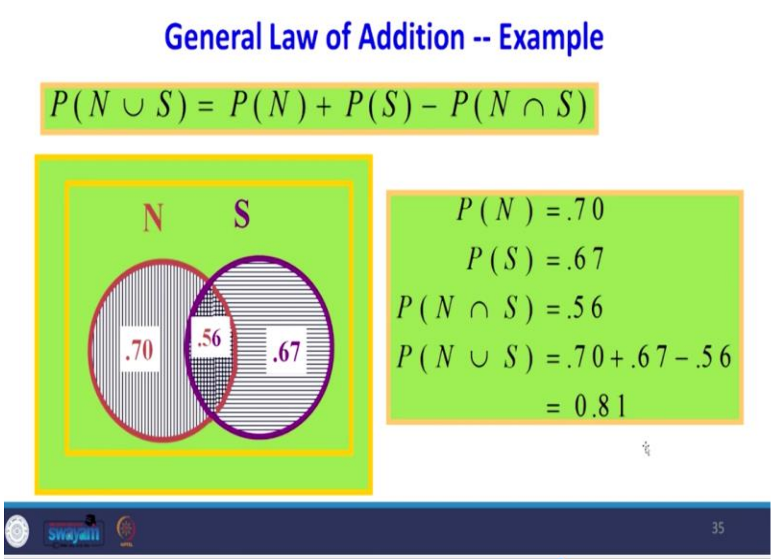
Suppose, if one of the survey respondents were randomly selected and asked what office design change would increase workers productivity, otherwise, what is the probability that this person would select reducing noise are it designed which is helpful for providing more storage space out of this that reduce out of 2 options.

(Refer Slide Time: 19:29)



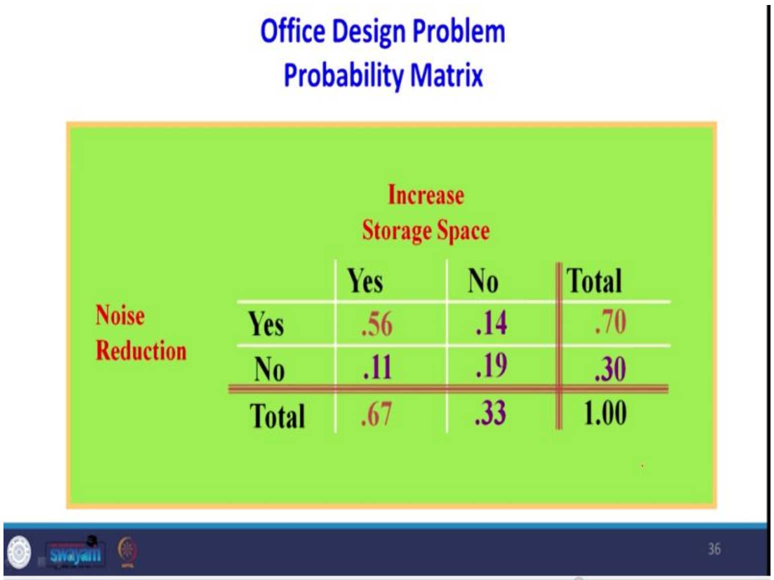
So, let N represent is the event reducing noise that means you are choosing that design yes represents the event more storage space, yes, that is the another options. The probability of person responding to N or S can be symbolized statistically as a union probability by using law of addition that is a P of N union S.

(Refer Slide Time: 19:56)



So, see that P of N union S is because we know they asked for the our the formula P of N union S is P of N + P of S - P of N intersection S. So, the P of N is 70% P of S is 67% those who have told S for both the designs is 0.56. So, when you substitute these values in the formula, so, we are getting 0.81 that is 81% of the people have told both the designs will increase the productivity

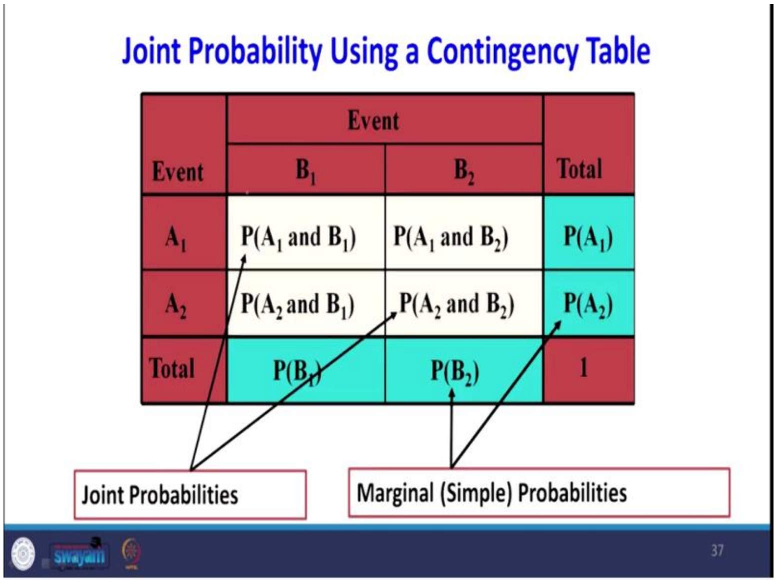
(Refer Slide Time: 20:33)



What you have solved with the help of Venn diagram in previous in the previous slide can be solved with the help of contingency table. This contingency table is so helpful just we have to make a table. For example, in rows I have taken noise reduction, the noise reduction say 70% people have told yes, so, remaining 30 might have told no. 70 30 in the column I have taken increasing storage space design in the 67% total they have told that increasing storage space would increase the productivity.

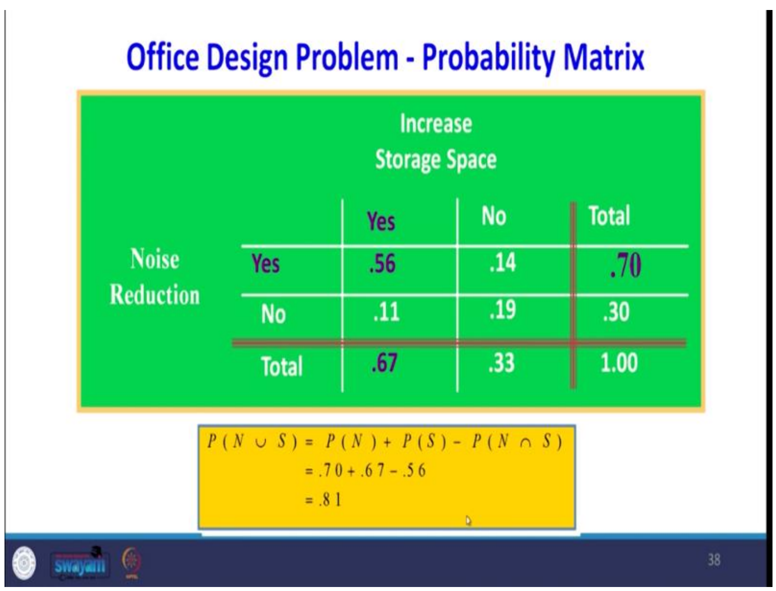
So remaining 33 might have told no. So, the 0.56 is intersection people have told both yes for both the design that is for noise reduction and increasing storage space. So, once if you know this 0.56 the remaining things you can simply you can subtract it 0.70 - 0.56 0.14 that is those who have told no to storage space and yes to noise reduction then if you subtract from 0.67 will get 0.11 0.33 - 0.14 we will get 0.19 so from this table we can read a lot of information’s.

(Refer Slide Time: 21:49)



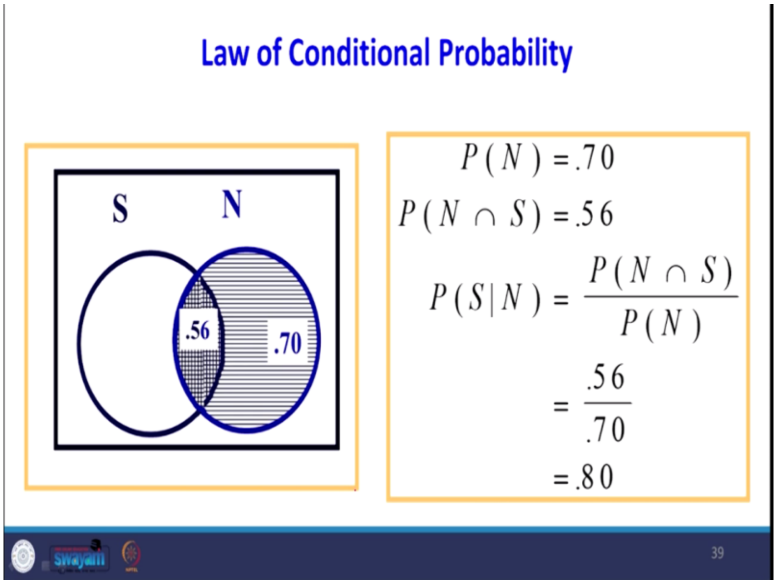
Whatever it has no this is for example, event A A1 A2 event B1 B2 this in the rows, you see that? It is in the columns whatever cell inside the cell this portion is called the joint probabilities P of A1 intersection B1 whatever it is the extreme side of the table see that this called marginal probabilities, this is a notation that traditionally they follow whatever the inside the cell it is a combination of both event that is the it is the joint probabilities whatever extremely and it is a marginal probability extreme side of the table.

(Refer Slide Time: 22:26)



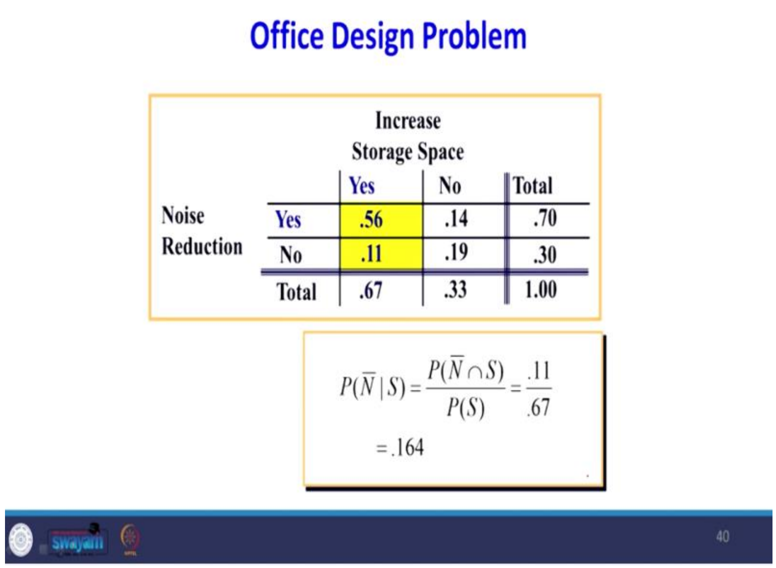
The same thing, suppose if you want to know the same answer with the help of this contingency table, we want to know how much percentage of the people who have agreed for both the design that is N and S. So, P of N + P of S - P of N intersection S. So this value directly we can read it from the table. So P of N is 0.7 0 + P of S is 0.67 - P of N intersection S is 0.56, so when you do that we getting 0.81.

(Refer Slide Time: 22:57)



Then we will go for conditional probability the probability of N is 0.7 0 those who are good both the design engineers is .56 suppose if you want to not pay off S \ N = P off N intersection is due to the by P of N so this value I will explain this conditional probability in detail later, but now you take this one, so, P of N intersection yes from the previous table you can find out the 0.56. You can look at the Venn diagram also, P of N is 0.7 and you substitute getting 0.8

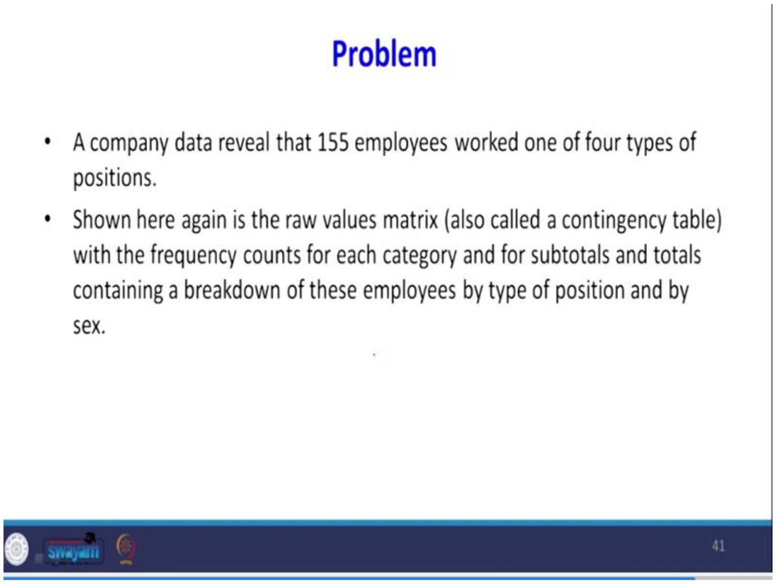
(Refer Slide Time: 23:37)



Noise Reduction	Increase Storage Space: Yes	No	Total
Yes	.56	.14	.70
No	.11	.19	.30
Total	.67	.33	1.00

The same office design problem you see that there is a another conditional probability P of N dash given Y that means, those who are told No to noise reduction, but they are agreed for storage design. So, this is the conditional probability. So, we have to multiply P of N dash intersection S divided by P of S N dash intersection S is this point because N no S this 0.11 divided by P of S 0.67 that is equal to 0.164.

(Refer Slide Time: 24:14)



We will take another small problem will explain the concept of probabilities with the help of the problem. A company data revealed that 155 employees worked on 4 types of positions the table is shown in the next slide raw value of matrix also called the contingency table with the frequency counts of each category and sub totals and totals containing breakdown of these employees by type of position and by sex.

(Refer Slide Time: 24:46)

Contingency Table

COMPANY HUMAN RESOURCE DATA

Type of Position		Sex: Male	Female	
	Managerial	8	3	11
	Professional	31	13	44
	Technical	52	17	69
	Clerical	9	22	31
		100	55	155

42

You see that look at this table in rows, the type of position they hold in their organization, whether they are working in managerial position, professional or technical or clerical in the column sex whether male or female, you see the intersection of managerial and male 8 that represents both that is more managerial working in a managerial position the same time they are male. So, that is our join values here the only count join counts the extreme right or the bottom of the table we are given the total counts.

(Refer Slide Time: 25:23)

Solution

- If an employee of the company is selected randomly, what is the probability that the employee is female or a professional worker?

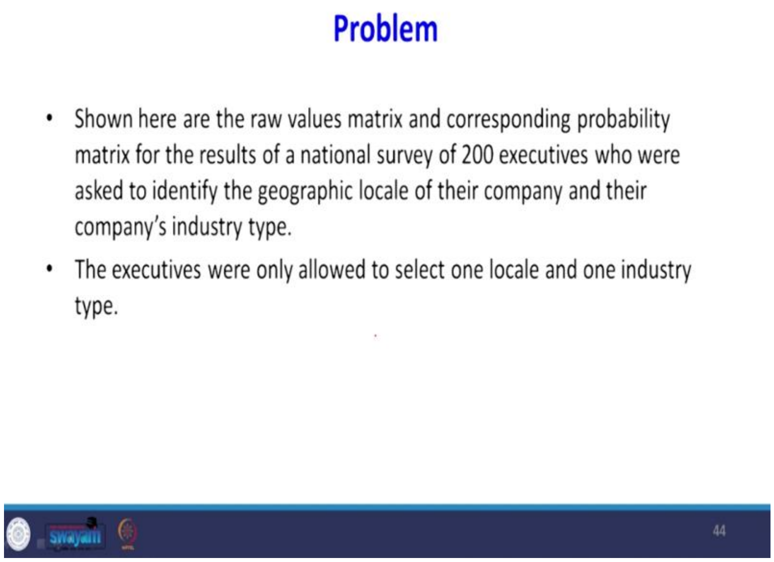
$$P(F \cup P) = P(F) + P(P) - P(F \cap P)$$

$$P(F \cup P) = .355 + .284 - .084 = .555.$$

43

Now, if an employee of the company is selected randomly, what is the probability that the employee is female or professional worker, so, what we have to do? So, we are going to find out P of F union P that is a P of F + P of P - P of F intersection P. So, P of F is when you go to the previous one P of F is so, when you divide 55 by 155 will get 0.335 P of P is when you divide 44 divided by 155 you will get 0.824 - P of F intersection P that is F intersection P 13 divided by 155 you will get 0.84 that equal to 0.55. There is another problem.

(Refer Slide Time: 26:15)



Shown here are the raw value matrix and corresponding probability matrix of the result of a national survey of 200 executives who were asked to identify geographic location of their company and their company's industry type. The executives were only allowed to select one location and one industry type, because it is not possible same person working different location different industries, because one person can work only one type of industry will conclude that in this session.

We have seen different types of probability how to assign probability and different counting rules may mn rules with replacement, without replacement. And different terms, which you are frequently we are going to use in this course that is the event, join probability, marginal probability and so on, then you have taken one sample problem with the help of sample problem, we have seen how to find out union of 2 events that is a joint probability P of A union B, then intersection P of A intersection B.

Then how to find out the marginal probability, then how to find out the conditional probability, but that will close and continue with the next lecture. Thank you very much